Applied regression analysis

Andrew Li

Spring 2021

Contents

Chapter 1

Preamble

1.1 Caution

This book was made for studying purposes only. I will be adding notes from other sources and I may leave out some topics from the course. Also, I will not be purchasing or reading the supplemental text. As such, this is not a faithful representation of the course.

1.2 Acknowledgments

This course was taught by Dr. Jason Rights!

1.3 Course goals

There are three primary goals of this course. The first goal is to provide students with sound foundational knowledge in the theory and concepts of linear regression analysis. The second goal is to develop an ability to properly apply regression methods to empirical data, including making informed decisions about analytic strategies and understanding how to report results. The third goal is to be able to critically evaluate the use of linear regression methods in research literature and the news. This course requires the successful completion of PSYC 218.

1.4 Readings

- Supplemental textbook (not required): Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences.* (3rd edition). Hillsdale, NJ: Erlbaum.

1.5 Course content

Note: Bolded weeks indicates that a problem set is due or there is a midterm.

Week	Date (2021)	Lecture topics	Readings	HW/MT
1	1/11 - 1/15	1. Orientation; Review of the Pearson correlation2. Factor affecting the size of correlation	1. (Ch. 1, 2.1 - 2.2) 2. (Ch. 2.3: 2.10)	none
2	1/18 - 1/22	1. Simple linear regression 2. Inferences for SLR	1. (Ch. 2.4 - 2.7) 2. (Ch. 2.8)	none
3	1/25 - 1/29	1. Multiple linear regression (MLR) with 2 IVs	(Ch. 3.1 - 3.2)	HW 1
4	2/1 - 2/5	1. MLR with k IVs	(Ch. 3.5)	none
5	2/8 - 2/12	*Lecture catch-up*	none	MT1
6	2/15 - 2/19	*Spring Break*	none	none
7	2/22 - 2/26	1. Power analysis for MLR 2. Hierarchical regression	1. (Ch. 3.7) 2. (Ch. 5.3)	HW2
8	3/1 - 3/5	1. Interactions in MLR	(Ch. 7.1 -7.4)	none
9	3/8 - 3/12	1. Assumptions in MLR: Definitions and testing	(Ch. 4.1 - 4.5)	HW3
10	3/15 - 3/19	1. Categorical IVs: Dummy coding 2. Categorial IVs: Effect coding	1. (Ch. 8.1 - 8.2) 2. (Ch. 8.3)	HW4
11	3/22 - 3/26	1. Nonlinear regression	1. (Ch. 6.1 - 6.2)	none
12	3/29 - 4/2	1. Logistic regression	1. (Ch. 13.2)	MT2
13	4/5 - 4/9	1. Multivariate models, mediation	none	none
14	4/12 - 4/14	*Lecture catch-up*	none	HW5

Chapter 2

Review of Pearson correlation

2.1 Consider two variables separately

- Suppose you are given two variables, X_1 and Y
- Each measured on an *interval* or *ratio* scale (for now).
  - **Interval scale:** equal intervals between scale points but has an arbitrary 0; permits +, - operations; An example would be the SAT score.
    - For example, we can add/subtract in a meaningful way but it doesn't make sense to multiple or divide SAT score because it assumes the 0 would be meaningful (the lowest possible SAT score is 400).
  - **Ratio scale:** equal intervals and a true 0 (which denotes absnce of construct); permits +, -, *, / operations; for example: Age.
    - Multiplying and dividing makes sense in this case because a 4 year old is twice as old as a two year old.
- With a sample size of n
- Scores for individual i are X_{1i} and Y_i.

Example data set:

- X_1 is the average monthly temperature baby first tries tto crawl (F)
- Y is baby's age of first crawl (in weeks)
- $n = 12$

observation	X_1 (temp)	Y (weeks)
1	66	29.84
2	73	30.52

observation	X_1 (temp)	Y (weeks)
3	72	29.7
4	63	31.84
5	52	28.58
6	39	31.44
7	33	33.64
8	30	32.82
9	33	33.83
10	37	33.355
11	48	33.38
12	57	32.32

2.1.1 Descriptive statistics

First, we consider the variables separately, looking at descriptive statistics for each variable. **Note:** When estimating a piopulation quantity, use $n-1$ to account for the loss of 1 degree freedom.

Mean:

$$\bar{X_1} = \frac{\sum X_{1i}}{n} = 50.25 \bar{Y_1} = \frac{\sum Y_i}{n} = 31.77$$

Variance (unbiased):

$$sd_1^2 = \frac{\sum (X_{1i} - \bar{X_1})^2}{n-1} = 251.11 sd_Y^2 = \frac{\sum (Y_i - \bar{Y})^2}{n-1} = 3.10$$

Variance (biased):

$$SD_1^2 = \frac{\sum (X_{1i} - \bar{X_1})^2}{n} = 230.19 SD_Y^2 = \frac{\sum (Y_i - \bar{Y})^2}{n} = 2.84$$

Standard deviation:

$$sd_1 = \sqrt{sd_1^2} = 15.85 sd_Y = \sqrt{sd_Y^2} = 1.76 SD_1 = \sqrt{SD_1^2} = 15.17 SD_Y = \sqrt{SD_Y^2} = 1.69$$

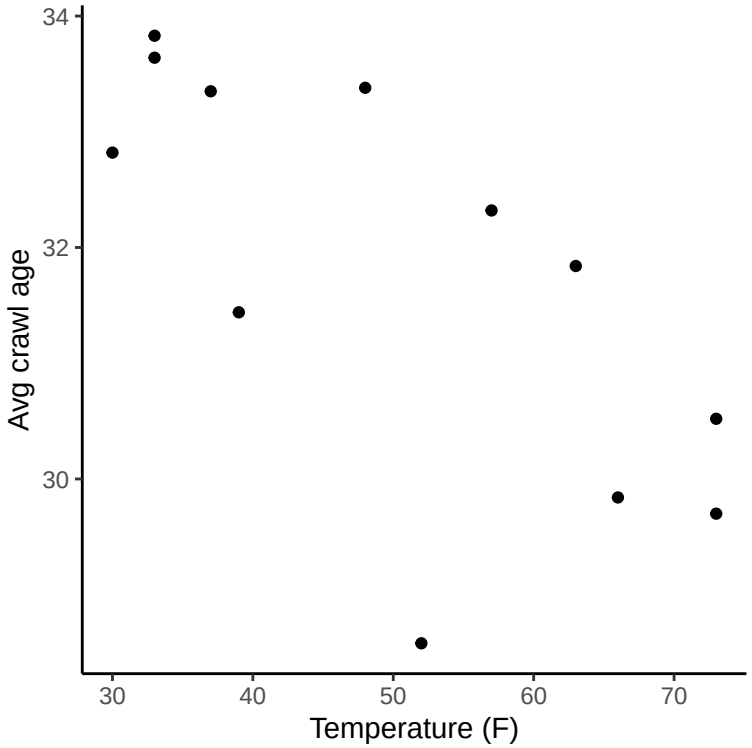
2.2 Relationnship between two variables

Now, let's consider the relationship *between* the two variables. The **hypothesis** is that infants take longer to learn to crawl in cold weather.

2.2.1 Graphical representation

We can graphically represent the relationship between two variables using a scatterlot. In a scatterplot, each axis represents one variable. Each observations are represented as points in space, with coordinates of pint corresponding to

scores on X_1 and Y. Here we are focusing on linear relationships. Scatterlot can provide some informal/qualitative information about the presences of a linear relationship.



However, it would be convenient to also have a *quantitative* measure summarizing the degree or strength of the linear relationship.

2.2.2 Covariance

One possible quantitative measure of linear association we could use is **covariance**. * Covariance between $X-1$ and Y:

$$c_{Y1} = \frac{\sum (X_{1i} - \bar{X_1})(Y_i - \bar{Y})}{n-1} = -19.53$$

- However, the covariance is *not* independent of the scales of X_1 and Y. Thus, its magnitude as affected by both the strength of linear relationship *and* the scales of the variables.
  - So if Y has variance of 1,000,000 and X_1 has a variance of10, it is hard to know what covariance means because of the different scales
- Consequently, the covariance is *not* bounded by -1 and +1.

2.2.3 Pearson correlation coefficient

So we would like a quantitative measure summarizing the degree or strength of the linear relationship. *And* we would like that measure to *not* change with ann arbitrary change in units of the variables…

We can eliminate scale issues for X_1 and Y by dividing by the standard deviations of both variables, yielding the **Peaarson correlation coefficient:**

$$r_{Y1} = \frac{\sum (X_{1i} - \bar{X_1})(Y_i - \bar{Y})}{(n-1)sd_1 sd_Y}$$

or, equivalently (c_{Y1}) was defined as the covariance:

$$r_{Y1} = \frac{c_{Y1}}{sd_1 sd_Y}$$

- The correlation is independent of scales of measurements because it uses standardized scores(z_1 and z_Y) which are not altered by linear transformation of raw scores (X_1 and Y).

$$r_{Y1} = \frac{\sum (X_{1i} - \bar{X_1})(Y_i - \bar{Y})}{(n-1)sd_1 sd_Y} = r_{Y1} = \frac{\sum z_{1i} z_{Yi}}{n-1}$$

- where z_{1i} and z_{Yi} are standardized scores (z-scores).

$$z_{1i} = \frac{(X_{1i} - \bar{X_1})}{sd_1} z_{Yi} = \frac{(Y_i - \bar{Y})}{sd_Y}$$

- Properties of standardized scores (z-scores)
  - Means are equal to 0
  - Variance always equal to 1
- You can find the standardized score in R usinf the `scale()` function:

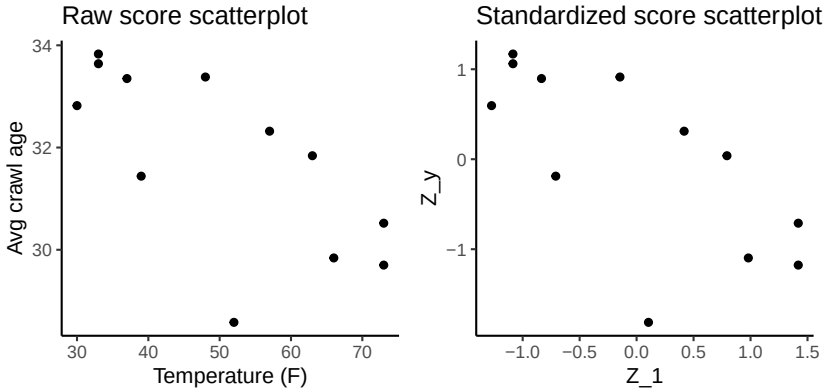
```
scale(data_crawl$temp)
```

- Standardized sscores(z-scores) represent relative standing with respect to the mean.
- Standardizing varibales does not change the ranking of scores
  - As you can see from the table below, the rank of z-scores are the same as the rank of raw scores.
  - Raw scores that were above/below the mean corresponds with z-scores above/below 0.

```
##   temp weeks        Z_1         Z_y       z1_zy
## 1   66 29.84  0.9807934 -1.09717461 -1.07610160
## 2   73 30.52  1.4190202 -0.71093886 -1.00883661
```

```
## 3   73 29.70  1.4190202 -1.17669374 -1.66975220
## 4   63 31.84  0.7929819  0.03881291  0.03077793
## 5   52 28.58  0.1043397 -1.81284675 -0.18915193
## 6   39 31.44 -0.7095101 -0.18838460  0.13366078
```

- As well, points in standardized scores scatterplot retain the same relation to each other as in raw scores scatterplot. The underlying relationship remains the same.



- Now let's calculate the Pearson correlation using the `data_crawl$z1_zy` column.

$$r_{Y1} = \frac{\sum z_{1i} z_{Yi}}{n-1} = \frac{-7.71}{11} = -.70$$

- Let us consider how this measure behaves under different relationships of variables:
  - Negative relationships: multiplying (+)(-) scores yields a negative sum of products
    - On average, being positive on x meant they were negative on y , meaning that we can expect a lot of negative values in the sum of products, thus resulting in a negative correlation.
  - Positive relationship: multiplying (+)(+) or (-)(-) yields a positive sum of products
  - Zero relationship: yields zero sum of products
- correlations are bounded by +1 (perfect positive correlation) and -1 (perfect negative relationship) because the variance of a standardized variable is 1/-1.

2.3 Interpreting a Pearson correlation

- **Magnitude**: measures strength of linear relationship (larger absolute values are stronger)
- **Sign** indicates direction of linear relationship

- **Independent of scales** of measurement
  - This is because standardized scores (z_1 and Z_Y) are noot alterated by linear transfoormatino of raw scores (X_1 and Y)

Chapter 3

Factors affecting correlation coefficients

3.1 Overview

Observed values of correlation coefficients are affected by a number of factors to be examined here:

Factor	Consequences	Fix?
Restriction of range	Usually reduce r	"Correction for restriction of range"
Outliers	Reduce or increase r	Ideally, detect a priori
Nonlinearity	Misleading r	Ideally detect a priori (later lecture)
Dichotomization	Usually reduce r	Don't dichotomize (corrections avalible but not discussed here)

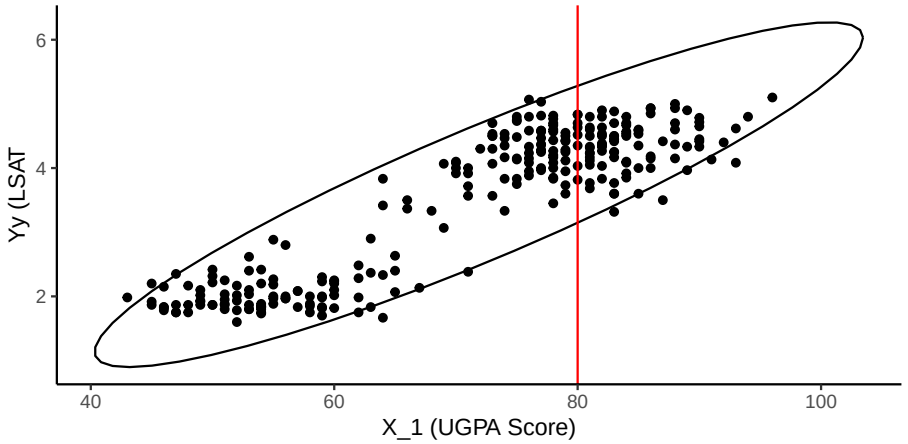
3.2 Restriction of range

Within a given law school, there is a near-zero correlation between undergraduate GPA and LSAT scores. This is an instance f restriction of range affecting the correlation.

Suppose we are interested in the realtionship between X_1 (UGPA) and Y (LSAT). However, individuals measured on X_1 are **selected** to be in the sample (admitted to law school) only if their score on X_1 is sufficiently high, exceeding some threshold.

In such situation, the range of X_1 is said to be *restricted*

It is useful ti distinguish between the correlation when selection had occured vs. the correlation if no selection had occurred. The difference can be seen in the following representation of a scatterplot:



If no selection occurs, the relationship between X_1 and Y is represented by the full ellipse. If selection occursm the relationship is represented by only the section of the ellipse beyoond the selection threshold on X_1. It is evident that the correlation would be greater without selection.

This phenomenon impacts interpretation of correlations in s selection situation. After selecting for X_1, the observed value of r_{Y1} may suggest a weaaker relationship then is actually present.

3.2.1 Correction for restriction of range

3.3 Outliers

3.4 Nonlinearity

3.5 Dichotomization

Chapter 4

Methods

We describe our methods in this chapter.

Chapter 5

Applications

Some *significant* applications are demonstrated in this chapter.

5.1 Example one

5.2 Example two

Chapter 6

Final Words

We have finished a nice book.